DOCUMENT RESUME

ED 464 946 TM 033 877

AUTHOR	Green, Donald Ross
TITLE	Interpreting the Results of Three Different Standard Setting Procedures.
PUB DATE	2001-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Cutting Scores; Elementary Secondary Education; Holistic Approach; Judges; *Standard Setting; *Teachers; Test Interpretation
IDENTIFIERS	Contrasting Groups Method; Jaeger Method; Kentucky

ABSTRACT

Different procedures for setting cut points on achievement test scales provide the standard-setting participants with different information to support the unique judgment task associated with each procedure. This study examined how participants in standard settings used the different information from three different procedures in Kentucky in 2000. Three cut points had to be established on the scales for each of six content areas in each of three grades for Kentucky's assessment system. The procedures were: (1) a contrasting groups procedure; (2) a holistic examination of student work on the tests as described by R. Jaeger and C. Mills (1998); and (3) a bookmark procedure (D. Lewis and others, 1998). Using Kentucky's five-step standard-setting process, the judges (teachers) established preliminary descriptors, conducted each of the three approaches, and then conducted a synthesis process to put together the results of the preceding steps. All told, 1,599 teachers participated in the exercises. Results show that each of the three standard setting procedures produced sets of cut scores for the various tests that can be said to be reasonable and defensible. Given the common backgrounds of the teachers in each of the three sets of judges and the carefully developed sets of descriptors of the categories, it is evident that much of the difference in cut scores stems from the different kinds of information the procedures provide. Of course, it is time consuming to use more than one procedure, and if only a single procedure is wanted, these data suggest the bookmark approach because of an apparently clearer relationship to instruction. The possibilities of the Jaeger-Mills process for helping schools improve could prove real, as many teachers wanted to see more full papers during the synthesis process. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 946

Interpreting the Results of Three Different Standard Setting Procedures

Donald Ross Green
CTB/McGraw-Hill

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033877

BEST COPY AVAILABLE

Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA, April 14, 2001.

Introduction

Different procedures for setting cut points on achievement test scales provide the standard setting participants with different information to support the unique judgment task associated with each procedure. This may explain in part, perhaps in large part, the long accepted and well-demonstrated fact that different standard setting processes produce different cut points (Jaeger, 1989).

The opportunity to observe how participants in a set of standard settings used the different information from three different procedures arose in Kentucky in 2000.

Kentucky's assessment system uses the results of tests in a number of grades and content areas as part of its school accountability system. The tests include both multiple choice items and open-ended items that are scaled together using item response theory. The open-ended items are scored with four point rubrics and are given twice the weight of the multiple-choice items. The system requires that each student's performance on which ever test they have taken be classified in one of four categories or performance levels: Novice, Apprentice, Proficient, or Distinguished (NAPD). Therefore three cut points had to be established on the scales for each of the six content areas in each of the three grades at which they were assessed.

Upon the recommendation of Kentucky's National Technical Advisory Panel for Assessment and Accountability it was decided to use three different procedures to determine these cut points

1. A contrasting groups procedure as outlined variously by Zieky & Livingston (1977), Poggio, Glassnap & Eros (1981), Livingston & Zieky (1982), and Jaeger, (1989).
2. A holistic examination of student work on the tests as described by Jaeger and Mills (1998).
3. A Bookmark procedure as described by Lewis et al (1998).

Using Kentucky's teachers as judges a five-step process was adopted:[1]

1. Establish preliminary descriptors or definitions of student performance at each of the NAPD levels for each of the grade/content areas. These descriptors were provided to each of the judges participating in each of the three standard setting processes.
2. Conduct a contrasting groups study in each of the 18 grade/content areas.
3. Conduct a Jaeger-Mills standard setting in each of the 18 grade/content areas.
4. Conduct a Bookmark standard setting in each of the 18 grade/content areas.
5. Conduct a synthesis process in which the results of the preceding steps were put together by teacher committees to lead to a final recommendation to the Kentucky State Board of Education.

Only major features of these processes will be described. The emphasis will placed on the information provided to the teachers (judges) by the three different

procedures (Steps 2, 3, and 4) and the ways in which the teachers made use of that information during those steps and at Step 5. A more detailed description of the steps can be found in the Kentucky Standard Setting Technical Report (CTB/McGraw-Hill, 2001) and in the report Standards Setting: Synthesis of Three Procedures—Procedures and Findings (Kentucky Department of Education, 2001).

The Five Steps

Step 1

Descriptors were developed for each grade/content area by four experienced Kentucky teachers. These groups worked both horizontally across content and vertically within content. Drafts were made at a two day meeting, posted on the Kentucky Department of Education website for a month for comment, refined at a further meeting of the drafting committee, and then approved as preliminary descriptors by the Kentucky State Board of Education. These descriptors were provided to the teachers participating in each of the three subsequent cut point settings.

Step 2

For the contrasting groups procedure a sample of approximately 50 schools for each grade/content area assessed by the Kentucky Core Content Test (KCCT) was

[1] The sixth and final step was consideration of the recommendations resulting from the Step 5 synthesis by the Kentucky State Board of Education. After study they did accept the descriptions and associated cut points as recommended.

selected by KDE. Each building principal selected two teachers in the relevant grade/content area. Shortly before the administration of the KCCT and after studying the descriptors, the teachers categorized all their students based on their classroom work as N, A, P, D or between one of the category pairs. An "insufficient knowledge" category was also available. The mean scale score of the students classified as N, A, P, or D was determined and the midpoint between the means of adjacent categories became the recommended cut point.

Step 3

The Jaeger-Mills procedures followed closely those outlined by Jaeger and Mills (1998). The basic process asked teachers to examine a number of student papers showing all the student's work on the test and then classify that work as high, medium, or low in one of the N, A, P, D categories. The design called for 18 teachers in each of the 18 grade content areas to work in groups led by curriculum experts provided by KDE. Craig Mills conducted the initial training for these group facilitators.

The 18 teachers in each content area were divided into two or three small groups depending on whether they were going to review work from two forms or three. After taking one form of the test, the teachers in each small group restudied the Kentucky Core Content and the descriptors from Step1 for their area in order to augment (e.g., add detail and/or delete unmeasured content) the descriptors to fit the test forms taken by the students whose work they would be evaluating.

Next each group examined a sample set of three papers and discussed how they classified them and how the descriptors were used to do this. Finally, each judge, working independently, classified 60 to 68 papers into the twelve categories using an iterative procedure. The performance of a student on the multiple choice items was examined first and then the responses to the open ended items were read and evaluated.

In four of the six content areas there were 24 multiple choice items and six four point open ended items. In the other two areas there were eight multiple choice items and two open ended items. The teachers were reminded that in obtaining a student's final score the open ended items were weighted twice as much as the multiple choice items.

Step 4

The Bookmark procedures used in Step 4 closely followed those outlined by Lewis et al (1998). In a Bookmark standard setting the judges are given an Ordered Item Booklet (OIB) in which each item in the test is presented in order of difficulty. Typically, and as in this case, the ordering is based on the item location derived from a calibration using two-and three-parameter models. The data came from the results of the statewide administration of the test. There is one item on each page and each score point from the polytomous items is treated as a separate item; thus the constructed response and multiple choice items are interspersed throughout the OIB. Items from more than one form can be

included if they are on a common scale. In this step, to more fully represent the domain being assessed, the items from two or three forms were used.

Working in three small groups of six to eight participants, the judges go through the OIB examining the items one by one, in order of difficulty, discussing what each item measures and why the item under consideration was harder than the preceding item. To facilitate this process an "item map" is provided. The item map lists one item per row with the item's scale location, the answer key for the MC items or the particular score point for the open-ended items. A typical time for this task was about three hours.

When the group completed their discussion of the entire set of items, each judge selected the point in the OIB such that, in their opinion, success on the items prior to that point constituted adequate evidence for classifying students in a given category. For example a judge might decide that if a student had demonstrated mastery of the content represented by items 1 to 50, he or she could be considered Proficient. The judge would then place a bookmark after page 50 of the OIB and the scale score location of that item would be the proposed cut point. The choices of bookmark locations were made independently in the first round of judgments. Once the choices were made the group discussed their disagreements and then made another choice. Finally all three groups met together and again discussed their disagreements before each judge made a third and final set of bookmark choices. The medians of the third choices of all the teachers in each grade/content area became the recommended cut points from the Bookmark standard setting.

Step 5

To produce final recommendations to the Kentucky State Board of Education, committees of Kentucky teachers spent two days studying the materials and recommendations of the prior committees in their grade/content area. Each Step 5 committee had between six and eight members. These committees contained participants from each of the preceding four steps, and also included someone from the adjacent grade level or levels.

The Step 5 committees were given:

- Draft Instructional Summaries prepared by KDE staff reflecting the results from each of the preceding steps.
- The results of the preceding standard setting committees in their grade content area.
- he Ordered Item Book and the item map used by the Bookmark committee.
- Sample student response booklets with scores close to the cut points chosen by the Jaeger-Mills committee.

These committees were asked to review the relevant test forms and all the materials, then, in the light of their view of the instructional consequences, select the results from one of the three methods as a starting point and either confirm or modify some or all of those three cut points. They were requested to stay within the range of scale scores defined by the three procedures. After making these preliminary decisions and writing

out their reasons (mostly done as a committee enterprise) they were given, in succession, information about:

- The impact: (the percent of students who took the 2000 test falling in each of the four categories)
- The cut points set by their colleagues in the other content areas at their level (elementary, middle school or high school) including impact. A chance to discuss these with their colleagues was provided.
- The cut points set by their colleagues in their content area at the other levels. Again a chance to discuss these with each other was provided.

After each of these sessions the committees met again and reconsidered their preliminary decisions. While some changes resulted, in almost all cases the changes made were minor even though many of the discussions within the committees were lengthy and in some cases heated; nevertheless substantial consensus did seem apparent in each of the 18 committees.

The participants

The judges were all Kentucky teachers. All of them had substantial experience teaching in Kentucky in their grade/content area. Table 1 shows the number of teachers and their years of experience in Kentucky. There do not appear to be any obvious differences in background experience and training between the participants in the three standard settings. Obviously they all had participated in the Kiris system and were fully

aware of the consequences for themselves and their school of the recommendations they were making about the cut points.

The Different kinds of information

Tables 2 and 3 summarize the information available to the standard setting judges at the various steps. It can be seen that all the participants had large amounts of information and that much of it was common to all (Table 2).

As Table 3 indicates, the Step 5 teachers had much, but not all, of the information available to the participants in both Steps 3 and 4, the Jaeger-Mills and Bookmark procedures. Furthermore they all had had the experience of participating in one of the preceding steps.

The teachers in the Contrasting Groups step had more information about the students they were rating than did the Jaeger-Mills participants. However they had not seen the students' work on the test and this difference is reflected in the large differences in the correlations between obtained score and ratings given by the two groups of judges illustrated in Table 4. An interesting phenomenon is the view of the Contrasting Groups teachers to the effect that there are more students properly labeled Distinguished than identified by either of the other two methods as shown by their preference for a relatively low P/D cut (Table 5). This was interpreted by a number of the Step 5 participants in several of the content areas as a limitation of the test in their grade/content area.

Unlike the Jaeger-Mills judges, the Step 5 judges did know the total scale score of the student work they examined; however they saw only about a third as many papers and a number of them stated a wish for more papers to examine, Furthermore staff observers report that in a number of cases the Step 5 judges found this scale information somewhat confusing because some papers had total scale scores that did not agree with their impression of the student's work.

A likely explanation based on unsystematic observation of the judges while they were discussing the papers is that the judges had difficulty in making use of the data from the multiple-choice items. One reason for this might be that when compared to the descriptions provided from Step 1, the teachers could more readily associate open-response kinds of evidence with these descriptions than multiple-choice kinds of evidence. Another element in this reaction may have come from their experience with KIRIS. The open-ended items use four point generic scoring rubrics, which during the KIRIS era were often translated directly into the four performance categories and were the only items that counted. Thus a score of 4 was interpreted as meaning a Distinguished performance, a 3 as Proficient and so forth regardless of the difficulty of the task even though item scores were never intended to be so interpreted.

Also some dismay was expressed by at least a few of the Step 5 participants about the variability in the quality of the responses to the open ended items within one student's work. The Step 5 science groups in particular were looking for papers that had responses

consistent with the performance level descriptions. Since the Step 3 judges did not discuss the papers they rated, these comments were not heard then (although the issue may have arisen during the work with the practice papers). These reactions may have played role in the wish expressed by a number of the Step 5 committees for more papers to examine.

To what appeared to be a lesser extent, some of the Bookmark participants also found the more directly visible differences in difficulty among the open-ended items at odds with their expectation or belief that the score points and the NAPD categories would or should have a one-to-one relationship. Although students can and do achieve a given score in a large number of different ways (i.e., different combinations of correct answers and score points), the Bookmark judges do not see this directly. Instead they talk about what a student should know and be able to do on the average.

The responses of the participants in the Jaeger-Mills and Bookmark sessions to the common items on the evaluation questionnaires were rather similar in most instances (Table 6), but clearly the open-ended items dominated the Jaeger-Mills judgments more than they did the Bookmark judgements. The Bookmark procedure appears to provide a more structured means of putting together the information from the multiple choice items and the open ended items than is the case with Jaeger-Mills procedure.

The Synthesis

In the first phase of the synthesis task the Step 5 committees were asked to choose one of the three sets of cut points as a starting point based on the degree they felt the

points chosen by the prior committee fit their instructional expectations. As can be seen in Table 7 the Bookmark cut was chosen as their starting point for discussion about 60% of the time and halfway between that and the Contrasting Groups cut another 20% of the time. Since the Bookmark defines the cuts in terms of content (the knowledge and skills required to answer the questions) and given the directions, it was probably natural that this was so. A related factor may have been the relative difficulty of the standards set by the three methods. Table 5 shows this comparison.

The Jaeger-Mills cut points were almost uniformly the highest (most difficult). Many teachers involved in the Jaeger-Mills process, and those using the sample of student response booklets in the Step 5 Synthesis indicated that individual student responses seemed somewhat inconsistent. That is, a student's paper might have generally had proficient-like attributes, but the response to a particular item would seem more like that of an apprentice student. The "inconsistency" that some judges struggled with is an outcome of the natural variation in performance that occurs in student work. This is not unexpected, and occurs in both directions, that is, performance on a particular item may be superior to overall performance for one student but inferior to it for another student. Because the classification was based on a holistic impression, the overall classification for such a paper would have tended to be proficient and because the variation occurred in both directions, it would not be expected to have an effect on the overall cuts score, which is based on means.

It may have been possible to select student response booklets with the desired distribution of scale scores and with consistent responses, however, they would not have represented typical student work and such a selection was not attempted.

The inconsistency in student performance may be the reason the cut points derived from these papers were rarely chosen as a starting point. The Jaeger-Mills cuts cores were uniformly farther from the final cut points selected by the Step 5 committees than those of the other two methods. Table 8 shows the scale score deviations of the cut points chosen by the Step 5 committees from those chosen by the three preceding committees. The synthesis process led the teachers to judge the Jaeger-Mills process as producing standards that were unreasonably difficult and the Bookmark results as closer to the appropriate standards.

An exception to this is the choice of the Contrasting Groups cut for identifying Distinguished performance. It is evident that in a number of the grade content areas the Step 5 teachers concurred in the opinion of the Step 2 participants. The Contrasting groups standard was chosen as the starting point for the Proficient/Distinguished cut by a third of the groups. Comments on the lack of opportunity to demonstrate distinction were heard frequently in many of the groups.

Conclusions

Each of the three standard setting procedures used produced sets of cut scores for the various tests that can be said to be reasonable and defensible. Given the common backgrounds of the teachers in each of the three sets of judges and the carefully developed common sets of descriptors of the categories, it is evident that much of the differences in cut scores stems from the different kinds of information the procedures provide.

Unless one can argue that the information produced by one of the procedures is definitive, it follows that the use of multiple approaches to standard setting is desirable. The view that getting different results from different procedures makes the them all seem invalid does not stand up once it can be seen that, although they are different, the bases of the judgments are each relevant.

Of course it is time consuming and therefore expensive to use more than one procedure. If forced to choose a single procedure these data suggest the Bookmark because of the apparently clearer relationship to instruction.

However the possibilities of the Jaeger-Mills process for helping schools improve could prove real. It is notable that during the Step 5 Synthesis many teachers wanted to see more full papers. It may be somewhat difficult to reconcile the total score with the holistic judgement in some cases but the advantage of considering real complete performances in classifying student test performance can be compelling.

One could argue that the information given by the Contrasting Groups procedure does not refer to performance on the test and therefore is not relevant. Although largely limited to the Proficient/Distinguished cut point the opinion of many of the Step 5 committees contradicts this view to say nothing of the opinion expressed by a majority of the thousand or so teachers in who felt that the test misclassified a number of their best students. While this view can be considered a problem of the test it also can be considered a criticism of the view that distinguished performance by a student on a test must be uniformly outstanding. In this sense it is a criticism of current thought about standards that does not recognize the inevitability of variability in human skills and performances.

Table 1
Experience of Kentucky Standard Setting Participants

Standard Setting Steps	N	Total Years Taught		Years Taught in Kentucky		Years Taught in Subject	
		M	SD	M	SD	M	SD
Descriptor Writing	85	16.9	7.5	16.1	7.5	11.8	7.8
Contrasting Groups	909	14.8	9.3	13.9	9.1	11.3	8.7
Jaeger-Mills	299	14.0	8.4	13.1	8.1	10.5	7.8
Bookmark	279	14.8	8.2	13.8	8.1	11.0	7.8
Step 5	127	14.7	8.5	13.7	8.2	11.6	8.1

Table 2

Summary of the Information Available to All Judges

Information
Extensive Knowledge of the Kentucky Core Content
Knowledge of the Step 1 Descriptors
Teaching Experience in Kentucky
Experience with KIRIS (the prior system)

Table 3
Summary of Information Available to the Various Sets of Judges

Information	Contrasting Groups	Jaeger-Mills	Bookmark	Step 5 Synthesis
Students' classroom performance	yes	no	no	no
Opinions of other teachers about student	yes	no	no	no
Sets of complete student responses to two (or three) forms of the test	no	yes	no	yes
The modification of step 1 descriptors deemed necessary to fit the forms studied	no	yes	no	yes
Relative difficulty of the items and score points	no	no	yes	yes
Judgements of other teachers about the cut points	no	no	yes	yes
Knowledge of the cut points set by other teachers in other content areas at the same level (primary, middle or secondary)	no	no	no	yes
Knowledge of the cut points set by other teachers for the same content area (across levels)	no	no	no	yes
Impact (percent of students falling within each category)	no	no	no	yes

Table 4
Some Correlations Between Ratings and Scale Scores

Grade	Content	Contrasting Groups	Ingegr-Mills
4	Reading	.580	.863
5	Math	.594	.803
7	Reading	.623	.841
8	Math	.585	.799
10	Reading	.598	.875
11	Math	.630	.817

Table 5
Percent of the Committees Selecting Relatively Low, Middle or High Cut Points

	Bookmark			Jaeger-Mills			Contrasting Groups		
Cut Score Achievement Level	Low	Middle	High	Low	Middle	High	Low	Middle	High
Novice/Apprentice	83%	14%	3%	0%	33%	67%	17%	53%	30%
Apprentice/Proficient	69%	31%	0%	3%	3%	95%	33%	67%	0%
Proficient/Distinguished	28%	67%	5%	0%	5%	95%	72%	28%	0%
Total	60%	37%	3%	1%	14%	85%	41%	49%	10%

Table 6

Committee Participants Qualitative Evaluation of The Process (by Percent)

Statement	Disagree		Neutral		Agree	
	Bookmark	Jaeger-Mills	Bookmark	Jaeger-Mills	Bookmark	Jaeger-Mills
I felt that the process was fair	2.2%	8.9%	7.0%	11.3%	90.6%	79.8%
The goals for this were clear	7.6%	18.9%	10.9%	13.0%	80.7%	67.7%
I gave more weight to the Constructed Response Items than the Multiple Choice Items	24.5%	4.7%	19.6%	1.7%	54.8%	93.0%
Reviewing the NAPD descriptors helped me decide how to rate student	1.1%	8.3%	7.4%	9.3%	89.8%	82.1%

Table 7
Frequency of Method Chosen for Starting Point

Method	N/A	A/P	P/D	Total
Bookmark	12	11	9	32
Split between Bookmark and Contrasting Groups	4	5	2	11
Contrasting Groups	2	1	6	9
Jaeger-Mills	0	1	1	2
Total	18	18	18	54

Table 8
Average Scale Score Deviations from the Step 5 Cut-Points

Level Boundary	Bookmark	Jaeger-Mills	Contrasting Groups
Novice/Apprentice	4.9	-20.3	-16.8
Apprentice/Proficient	1.7	-29.5	-7.3
Proficient/Distinguished	3.8	-30.2	12.1

References

CTB/McGraw-Hill. (2001). *Kentucky Standard Setting Technical Report.* Monterey, CA: Author.

Kentucky Department of Education. (2001). *Standards Setting: Synthesis of Three Procedures—Procedures and Findings.* Frankfort, KY: Author.

Jaeger, R. (1989). *Certification of Student Competence.* In R.L. Linn (Ed.), Educational Measurement (3rd ed., pp. 485-514). New York: Macmillan

Jaeger, R. & Mills, C. (1998). *An Integrated Judgment Procedure for Setting Standards on Complex Large-Scale Assessments.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 13-17, 1998.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., Patz, R. J. (1998). *The Bookmark Standard Setting Procedure: Methodology and Recent Implementations.* Paper presented at the 1998 annual meeting of the National Council on Measurement in Education.

Livingston, S.A., & Zieky, M.J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

Poggio, J.P., Glassnap, D. R. & Eros, D.S. (1981). *An Evaluation of Contrasting Groups Methods for Setting Standards.* Paper presented at the annual meeting of the American Educational Research Association, New York, New York, March 19-23, 1982.

Zieky, M.J., & Livingston , S.A. (1977). *Manual for Setting Standards on the Basics Skills Assessment Tests.* Princeton, NJ: Educational Testing Service.